SCALABLE LEARNING FOR RESTRICTED BOLTZMANN MACHINES

Elnaz Barshan, Paul Fieguth

Department of Systems Design Engineering
University of Waterloo
Waterloo, Canada

ABSTRACT

We propose Eigen-RBM, a scalable Restricted Boltzmann Machine (RBM) for visual recognition in which the number of free parameters to learn is independent of the image size. Eigen-RBM exploits the global structure of the image and does not impose any locality or translation-invariance assumption, and regularizes the network weights to be a linear combination of a set of predefined filters. We show that, compared to basic RBM, Eigen-RBM can achieve similar or better performance in both recognition and sample generation with significantly less training time.

Index Terms— Image Classification, Generative Models, Feature Extraction, Restricted Boltzmann Machines, Machine Learning

1. INTRODUCTION

Image understanding is a shared goal in all computer vision problems. This objective includes decomposing the image into a set of primitive components through region segmentation, region labeling and object recognition, and then modeling the interactions between the extracted primitives. However, due to large intra-class variations, extracting image primitives is highly challenging. Although images are given as a gridded set of pixels, in order to cope with large variations a high-level abstraction is required. Therefore, a key challenge is to bridge the gap between low-level pixel-representations and high-level abstract image descriptors.

The past decade was not successful in developing an effective abstraction mechanism. Researchers have tried to engineer hand-crafted descriptors to discriminate different components of the image [1, 2, 3, 4], where the best representative of this category is the Scale-Invariant Feature Transform (SIFT) [1]. Although successful, these discriminative models are domain-specific and require a large amount of labeled training data. On the other hand, the problem of visual recognition has some innate challenges, including occluded images and a lack of labeled data. Generative models address this issue by imposing additional constraints on the model parameters to perform well in generating images as well as discriminating them. That is, instead of hand-crafting image descriptors, features constituting a generative model can be used to regularize the parameters of a discriminative model. In this way, one can take advantage of the large amount of easily available *unlabeled* data while the generative aspects of the features can be investigated to deal with image occlusion. Considering these properties, we have witnessed a growth of interest in learning image descriptors in probabilistic generative frameworks [5, 6, 7, 8].

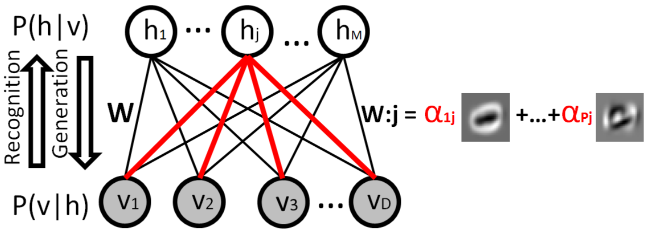


Fig. 1. Schematic diagram of Eigen-RBM in which filters (i.e. $w_{:j}$) are defined as linear combinations of a set of pre-defined filters.

Despite the merits of employing generative models in visual recognition, there are many unresolved obstacles. Of these, one of the most important issues is that of high computational complexity. Although considerable advances have been made [6, 9], running these algorithms requires special hardware and software support. In this paper, we focus on Restricted Boltzmann Machines (RBMs), the most widely-used generative model for visual recognition, and study the issue of computational complexity. We present Eigen-RBM, as an extension of RBM, which reduces the computational burden of RBM learning significantly. The key idea behind Eigen-RBM is to reduce the number of parameters by defining the network weights as linear combinations of a set of predefined filters.

The rest of this paper is organized as follows: Section 2 gives a brief overview, Section 3 introduces the proposed method and the performance of the proposed method is examined on a number of visual recognition tasks in Section 4.

2. BACKGROUND

2.1. Generative Models for Images

The basis of generative modeling approaches for image applications is the seminal work by Geman and Geman [10]. The heart of any generative approach is modeling the *prior* distribution of the data. However, due to the high-dimensionality and non-Gaussian statistics of natural images, defining a generic prior model is quite challenging. Researchers have evolved this model of [10] by investigating richer priors [11, 8, 12, 13].

We have witnessed a growth of attention [5, 6, 9] toward using *generative* models for visual recognition. The problem of visual recognition has some innate challenges, including a lack of labeled data and image occlusion, both of which can be addressed in a generative framework:

- Unsupervised Learning: Labeled data is costly to produce, however generative models are able to learn from unlabeled data.
- Occlusion: Investigating the generative property of a model, ambiguities in the raw sensory inputs can be resolved by means of inferring the missing pixels [6, 5].

2.2. Restricted Boltzmann Machines

Restricted Boltzman Machines (RBMs) [14] are the most widely used generative model for feature extraction in visual recognition [9, 15, 16, 17]. RBMs are bipartite undirected graphical models with a set of binary hidden units $\mathbf{h}$ and a set of visible units $\mathbf{v}$ (binary or real-valued) arranged in two layers. There are symmetric connections between these two layers represented by weight matrix W (see figure 1). In this structure, visible units correspond to input data (e.g., image pixels) and hidden units are the extracted abstract representations. In order to capture more abstract features, RBMs can be stacked to form deep architectures such as Deep Belief Networks [9] and Deep Boltzman Machines (DBMs) [18].

In an RBM, the probability of a configuration $(\mathbf{h}, \mathbf{v})$ is

$$P(\mathbf{h}, \mathbf{v}) = \frac{1}{Z} exp(-E(\mathbf{h}, \mathbf{v})) \tag{1}$$

where Z is the partition function, with energy $E(\mathbf{h}, \mathbf{v})$

$$\begin{aligned} E(\mathbf{h}, \mathbf{v}) = \frac{1}{2} \sum_{i \in vis} (v_i - c_i)^2 - \sum_{j \in hid} h_j b_j \\ - \sum_{i \in vis, j \in hid} v_i w_{ij} h_j \end{aligned} \tag{2}$$

where b and c are the biases of the hidden and visible units. Given the joint distribution $P(\mathbf{h}, \mathbf{v})$ (1), the probability that an RBM assigns to an input vector $\mathbf{v}$ can be obtained from the marginal $P(\mathbf{v})$. The parameters of an RBM can be optimized in a purely unsupervised manner, based on maximizing the likelihood of the training data. Since maximizing $P(\mathbf{v})$ with respect to the network weights does not have a closed-form solution, gradient ascent is applied.

One key disadvantage is the large number of parameters to be learned, thus the application of RBMs is limited to small-size images. Furthermore, having a large number of parameters is always a threat to the good generalization of a model. One way of tackling this problem is to reduce the number of parameters using a weight-sharing scheme [6, 5, 16]. Convolutional architectures [6] assume that the network weights (i.e., filters) are local and stationary, however, in practice translation invariance is frequently violated.

3. Eigen-RBM

The number of RBM parameters grows roughly quadratically with the size of the input image. Therefore, extending RBMs to high-resolution images is not computationally tractable or desirable, since having large numbers of parameters is a threat to good model generalization.

Suppose that $\mathcal{X} \in \mathbb{R}^D$ is a set of observed input variables and $\mathcal{Y} \in \mathbb{N}$ is a set of latent output variables drawn jointly from distribution $P(\mathcal{X}, \mathcal{Y})$. Given a set of unlabeled observed samples $X = \{x_1, x_2, \dots, x_N\}$ as N realizations of $\mathcal{X}$, we are looking for a weight matrix W which maps x to $\mathbf{h}$ such that the likelihood of the observations is maximized. We use $W_{:j}$ to denote the weight vector that connects all of the units of the visible layer to hidden unit h_j.

In order to scale RBMs to realistic-sized images, we propose that the weights $W_{:j}$ to be defined as linear combinations of a set of predefined filters (see figure 1). That is, given a filter bank $F = \{\mathbf{f}^1, \mathbf{f}^2, ..., \mathbf{f}^P\}$, we define weight vector (i.e., filter) $W_{:j}$ as

$$W_{:j} := \sum_{k=1}^{P} \alpha_{kj} \mathbf{f}^k \tag{3}$$

where the size of the filter bank is much smaller than the number of visible units (i.e., $P \ll D$). In this way, the number of parameters is independent of the image size and becomes instead a function of the size of the filter bank. In addition, the global structure of the image is exploited and no locality or translation invariance assumption is imposed.

Training an RBM consists of learning the weights α_{kj} of the filters in the filter bank. Performing gradient ascent on the log-likelihood of the training data, the update rule for coefficient α_{kj} is computed as

$$\begin{aligned} \frac{\partial logP(\mathbf{v})}{\partial \alpha_{kj}} &= \sum_{i \in vis} \frac{\partial logP(\mathbf{v})}{\partial w_{ij}} \frac{\partial w_{ij}}{\partial \alpha_{jk}} \\ &= \sum_{i \in vis} (< v_i h_j >_{data} - < v_i h_j >_{model}) f_i^k \end{aligned} \tag{4}$$

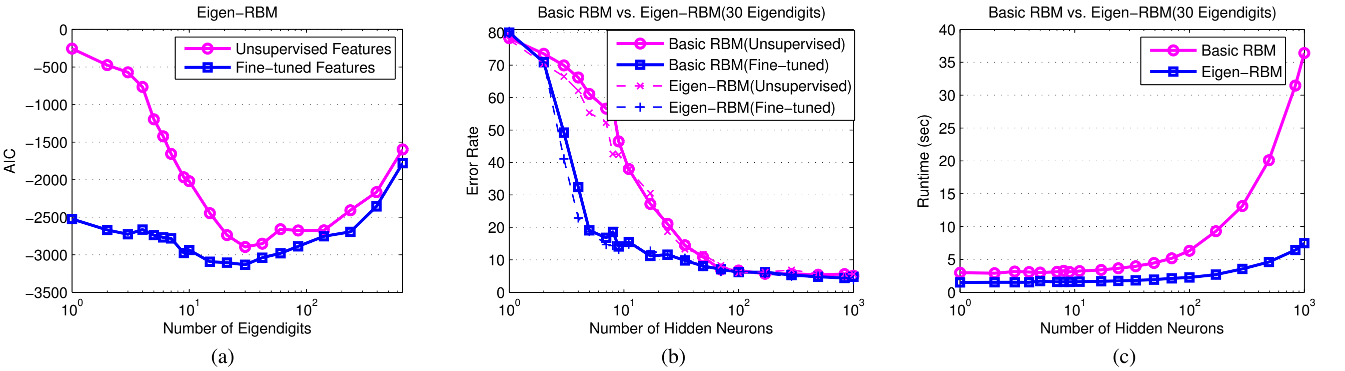


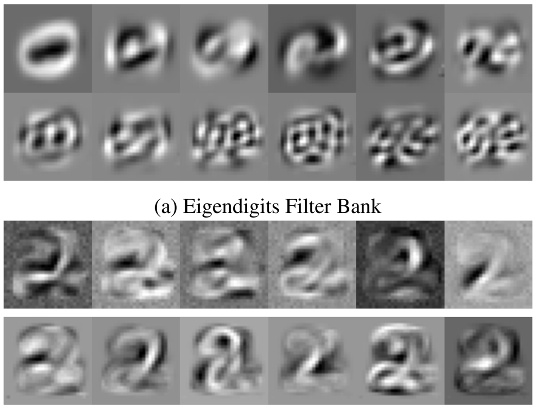
Fig. 2. (a) The Akaike Information Criterion (AIC) for Eigen-RBM on the Small MNIST data set. Note that for both unsupervised and fine-tuned features, the optimum number of Eigendigits, found by minimizing the AIC, are similar. (b) and (c) compare the basic RBM and Eigen-RBM in terms of classification error rate and running time. Observe that the Eigen-RBM has a similar error performance relative to basic RBM, but with much less training time.

The freedom of choosing the filter bank enables us to capture different aspects of the image. Since the filters are applied linearly to the image (rather than being convolved), a good choice would be a filter bank that captures global information at different levels of details. A simple choice for the filter bank could thus be the set of eigenvectors corresponding to large eigenvalues of the covariance matrix of the training data. Figure 3 (a) shows a sample of this filter bank for hand-written digits of MNIST data set [19]. The RBM with the weights defined as linear combinations of Eigen-filters is Eigen-RBM.

To assess whether Eigen-RBM can produce similar results to basic RBM, which has many more free parameters, both networks are trained with images of handwritten digit **2** taken from the MNIST data set. The filter bank in Eigen-RBM consists of the top 30 eigenvectors. As figure 3 (b) shows, with about one twenty-seventh the number of parameters, Eigen-RBM produces similar results to RBM. Although noise-reduction is not the objective, it is interesting to observe the reduction of noise in the corners of the learned Eigen-RBM filters; the Eigen-RBM filters are random combinations of Eigendigits, which are not noisy, whereas the learned filters for basic RBM are initialized with random values.

4. RESULTS AND DISCUSSION

We conducted a set of experiments on a small version of MNIST[1] data set of handwritten digits to study the effectiveness of Eigen-RBM compared to that of RBM. The feature learning procedure consists of two stages. In the first stage, the generative weights are learned in a purely unsupervised manner. In the next phase, the learned generative weights are fine-tuned using error backpropagation. The extracted representations are used to classify the test set using a one-nearest-neighbor classifier [20].



(a) Eigendigits Filter Bank

(b) Basic RBM (top) and Eigen-RBM (bottom)

Fig. 3. (a) In the filter bank, observe how the Eigendigits capture information at a variety of scales from coarse to fine. (b) The learned filters for single digit training data (digit 2) as computed by basic RBM and Eigen-RBM.

In order to determine the optimum number of Eigendigits for Eigen-RBM, we use the Akaike Information Criterion (AIC) [21], a model selection criterion rewarding goodness of fit while penalizing the number of free parameters. Figure 2 (a) presents the computed AIC for Eigen-RBM for different numbers of Eigendigits, illustrating that the required number of Eigendigits for both unsupervised and fine-tuned models are similar. Based on this criterion, in Eigen-RBM we use the top 30 Eigendigits of the data.

Figure 2 (b) and (c) propose a comparison between basic RBM and Eigen-RBM in terms of recognition rate and running time. To ensure a fair comparison, both RBM and Eigen-RBM are trained with the same number of epochs. This result demonstrates that with near one twenty-seventh the number of the parameters of basic RBM, Eigen-RBM performs similar to and even better than basic RBM. As illustrated in

[1] A subset of 600 and 100 samples from each digit class is chosen randomly for training and testing, respectively.

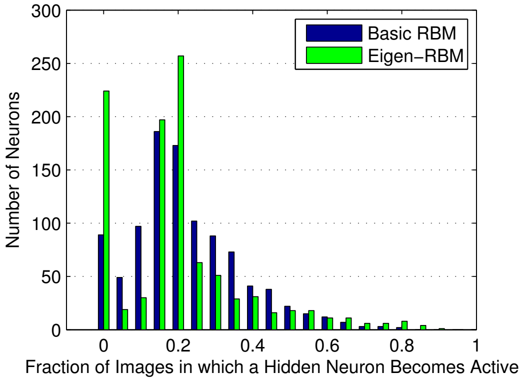


Fig. 4. A histogram of active hidden neurons on small MNIST data set, showing the number of hidden neurons that are active for some fraction of the images, such that the left-most bin shows the number of neurons that are active for none of the images (i.e., always off), allowing us to conclude that the Eigen-RBM representation is sparser than that of basic RBM.

figure 2 (c), for a large number of hidden neurons—which improves the recognition performance—the difference between the training time of basic RBM and Eigen-RBM becomes significant.

It is important to know that learning overcomplete representations[2], such as RBM features, run the risk of learning trivial solutions [6]. To avoid this problem, one can encourage the features to be sparse so that for each input only a small fraction of hidden neurons becomes active [22, 23]. Without imposing any sparsity penalty, a side-effect of the new weight-learning algorithm is to learn sparse representations. The histogram in Figure 4 shows how often the hidden neurons are active for the training images. Evidently, compared to basic RBM, in Eigen-RBM there are many more hidden neurons that are never active or active for a small fraction of time.

For the next step of this comparative study, we take a look at the mind of the network to see what the network believes in. For each class of digits on the complete MNIST data set, an RBM and an Eigen-RBM both with 50 hidden units are trained. After training, starting from a random binary input for the hidden layer, we run alternate Gibbs sampling for 500 iterations. Figure 6 shows the generated samples for each class at different Gibbs sampling iterations. As this figure shows, compared to basic RBM, Eigen-RBM, with much less free parameters, produces similar or better samples (e.g., digit 6, 7 and 8). In our last experiment, the average number of active hidden units is measured on 1000 samples drawn from each class of digits. As figure 5 illustrates, for all digit classes, Eigen-RBM generates similar or better samples with more sparse representations than that of basic RBM.

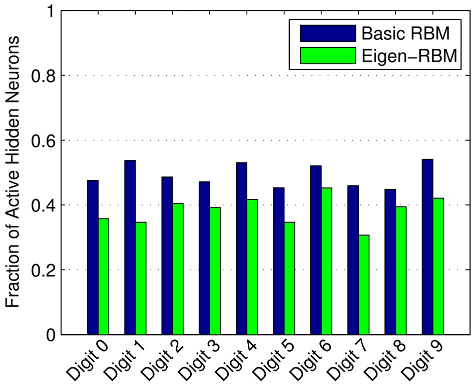
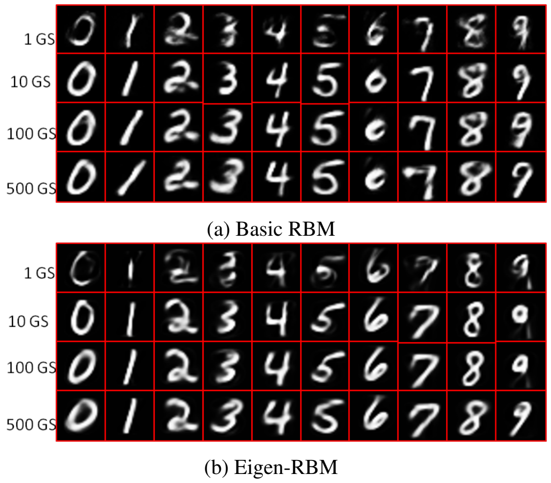


Fig. 5. Fraction of hidden units which are active for the generated samples of each class using basic RBM and Eigen-RBM. For all digit classes Eigen-RBM generates similar or better samples with more sparse representations than basic RBM.



(a) Basic RBM

(b) Eigen-RBM

Fig. 6. Generated samples using basic RBM and Eigen-RBM trained on single digit classes. The rows show the sample evolution as a function of Gibbs sampling iterations. Eigen-RBM, with fewer free parameters, produces similar or better samples (e.g., digits 6, 7 and 8).

5. CONCLUSION

This paper presented Eigen-RBM, with a scalable weight-learning algorithm in which the number of free parameters is independent of the image size. Compared to basic RBM, Eigen-RBM has similar or better performance in both recognition and sample generation with much less training time. Without imposing any sparsity regularization, the new weight learning algorithm leads to more sparse representations, the subject of future work.

[2] More sources of features than observations.

6. REFERENCES

[1] David G Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] Navneet Dalal and Bill Triggs, “Histograms of oriented gradients for human detection,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 886–893.

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, “Surf: Speeded up robust features,” in *Computer Vision–ECCV 2006*, pp. 404–417. Springer, 2006.

[4] Liefeng Bo, Xiaofeng Ren, and Dieter Fox, “Kernel descriptors for visual recognition.,” in *NIPS*, 2010, vol. 1, p. 3.

[5] M Ranzato, Joshua Susskind, Volodymyr Mnih, and Geoffrey Hinton, “On deep generative models with applications to recognition,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2857–2864.

[6] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng, “Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations,” in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 609–616.

[7] Alex Krizhevsky, Geoffrey E Hinton, et al., “Factored 3-way restricted boltzmann machines for modeling natural images,” in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 621–628.

[8] Stefan Roth and Michael J Black, “Fields of experts: A framework for learning image priors,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 2, pp. 860–867.

[9] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh, “A fast learning algorithm for deep belief nets,” *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[10] Stuart Geman and Donald Geman, “Stochastic relaxation, gibbs distributions, and the bayesian restoration of images,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, , no. 6, pp. 721–741, 1984.

[11] Max Welling, Simon Osindero, and Geoffrey E Hinton, “Learning sparse topographic representations with products of student-t distributions,” in *Advances in neural information processing systems*, 2002, pp. 1359–1366.

[12] Yair Weiss and William T Freeman, “What makes a good model of natural images?,” in *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[13] MarcAurelio Ranzato, Volodymyr Mnih, and Geoffrey E Hinton, “Generating more realistic images using gated mrf’s,” in *Advances in Neural Information Processing Systems*, 2010, pp. 2002–2010.

[14] Geoffrey E Hinton, “Training products of experts by minimizing contrastive divergence,” *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[15] Uwe Schmidt and Stefan Roth, “Learning rotation-aware features: From invariant priors to equivariant descriptors,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2050–2057.

[16] Gary B Huang, Honglak Lee, and Erik Learned-Miller, “Learning hierarchical representations for face verification with convolutional deep belief networks,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2518–2525.

[17] Yichuan Tang, Ruslan Salakhutdinov, and Geoffrey Hinton, “Robust boltzmann machines for recognition and denoising,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2264–2271.

[18] Ruslan Salakhutdinov and Geoffrey E Hinton, “Deep boltzmann machines,” in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 448–455.

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[20] C.M. Bishop et al., *Pattern recognition and machine learning*, vol. 4, springer New York, 2006.

[21] Hirotugu Akaike, “A new look at the statistical model identification,” *Automatic Control, IEEE Transactions on*, vol. 19, no. 6, pp. 716–723, 1974.

[22] Honglak Lee, Chaitanya Ekanadham, and Andrew Ng, “Sparse deep belief net model for visual area v2,” in *Advances in neural information processing systems*, 2007, pp. 873–880.

[23] Kihyuk Sohn, Dae Yon Jung, Honglak Lee, and Alfred O Hero, “Efficient learning of sparse, distributed, convolutional feature representations for object recognition,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2643–2650.